# Multimedia Quality of Experience for Target Recognition Applications

Mikołaj Leszczuk, Lucjan Janowski; AGH

# Presentation Plan

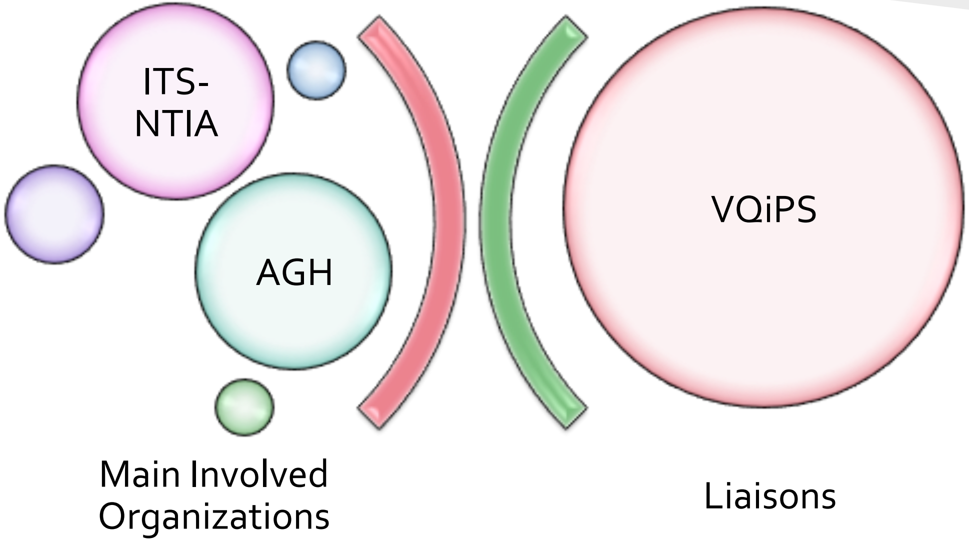
- Introduction
- Target Recognition Video (TRV) for the purpose of use
- Methods for subjective evaluation of TRV
  - Source signal
  - Testing methods, including multiple/single-choice
  - Subjects
  - Instructing and training subjects
  - Conditions for testing
  - Statistical analysis and reporting
- Summary - standardization

# *Introduction*

# VQEG’s Subproject: Quality Assessment for Recognition Tasks (QART)

- **Mission**
  - *“To study effects of resolution, compression and network effects on quality of video used for recognition tasks”*
- **Goals**
  - To perform series of tests to study effects and interactions of
    - Compression
    - Scene characteristics
  - To test existing or develop new objective measurements that will predict results of subjective tests of visual intelligibility
  - Propose subjective test methodology for recognition tasks

# Main Organizations and Liaisons



# *TRV for the Purpose of Use*

# Task Recognition Specificity (1/2)

- In many applications video quality not as important as ability to accomplish specific task for which video was created
- Typical examples of such **TRV**:
  - Video surveillance systems
  - Telemedicine / remote medical consultation / diagnosis system
  - Fire safety
  - Backup camera installed in car helping to park
- Quality tests needed
- General idea behind quality tests for TRV: to find threshold at which task can be achieved with certain probability or accuracy

# Task Recognition Specificity (2/2)

- Therefore, instead of quality evaluation, subjective experiment focused on task performance measurement
- For example, test might be measuring probability of:
  - For a video surveillance – recognition of license plate numbers
  - For telemedicine / remote diagnosis – correct diagnosis
  - For fire safety – fire detection
  - For backup camera – parking the car

# But What Evaluation Procedure to be Taken?

- Method of selection of source TRV, from which test TRV (with degraded quality) arise
- Subjective testing methods and general manner of conducting the psychophysical experiment
- Method of selecting a group of subjects in the psychophysical experiment
- Instructing and training of subject before the start of the experiment
- Conditions in which the test will be carried out
- Methods of statistical analysis and presentation of results

# *Methods for Subjective Evaluation of TRV*

# ITU-T Recommendation P.912

- Problems of quality measurements for TRV **partially standardized** in ITU Recommendation P.912
- Title: “Subjective Video Quality Assessment Methods for Recognition Tasks”
- Published: 2008
- Introducing:
  - Basic definitions
  - Methods of testing
  - Psycho-physical experiments

# Source Signal (1/2)

**Section 5 of P.912:**

*Test sequences should follow the general principles stated in [ANSI T1.801.01-1995] and [ITU-T P.910], which specify scenes that should be consistent with the transmission service under test, and should span the full range of spatial and temporal information. It is critical for the nature of these evaluations that the stimuli used actually reflect the true operational parameters of the conditions under which the video material is collected about, and cover the entire range of possible scenarios for the application area identifying that one is.*

# Source Signal (2/2)

- High data diversity, examples:
  - X-ray diagnosis of bone fractures
  - Accuracy of license plate recognition
- For some cases, data availability very limited
- Unjustified attempts to extrapolate scope
- Explicit instruction in P.912 needed, so proposed

# Testing Methods (1/2)

**Section 6 of P.912:**

*The application of TRV is directly related to the ability of the user that recognizes targets at increasing levels of detail. These levels are referred this as Discrimination Classes (DC). When determining the DC for particular scenarios, they must consider that for a set distance from the camera to the object of interest, the DC directly correlates video is decreasing resolution of the target, and therefore the object is represented by fewer cycles per degree of resolution. Fewer cycles per degree of resolution also means that the object subtends less of the information content of the video, making identification of the target more difficult.*

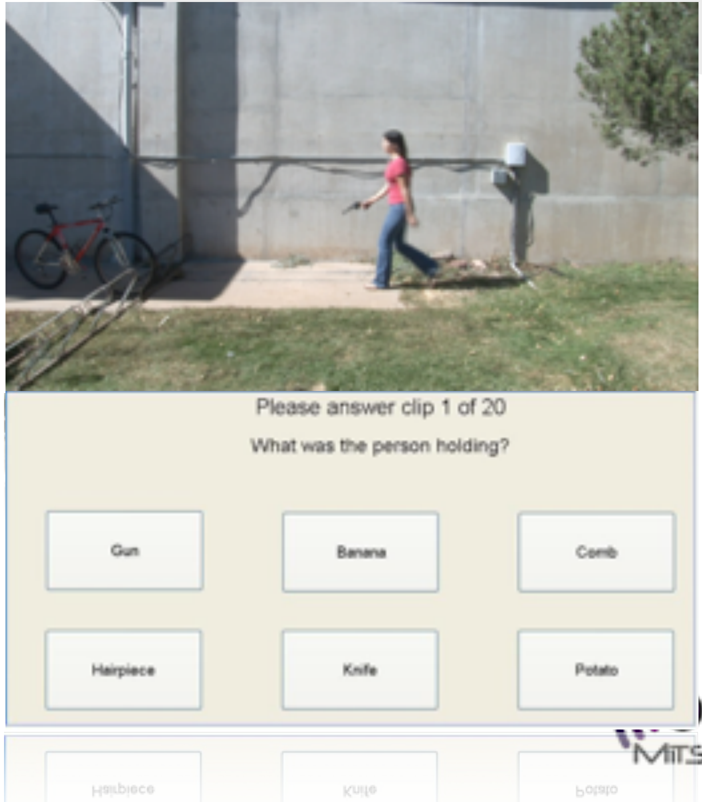
# Testing Methods (2/2)

- Not easy to understand relationship between parameters such as:
  - Number of Cycles-Per-Degree (CPD)
  - Resolution of the object, and
  - Distance between camera and object
- CPD - key parameter is CPD, affected by:
  - Resolution of object, and
  - Distance between camera and object (potentially)
- Changes involving ordering relationship between parameters proposed

# Multiple-Choice Method (1/4)

**Section 6.1 of P.912:**

*This method is appropriate for all DC levels and target categories (human, object and alphanumeric). For this method, the video is shown above a letter of verbal labels representing the possible answers. After presenting the video, the viewers must choose the label closest to what they recognized in the clip. The use of fixed multiple choices eliminates any possible ambiguity that could accommodate arise from open questions, and allows for more accurate measurements.*

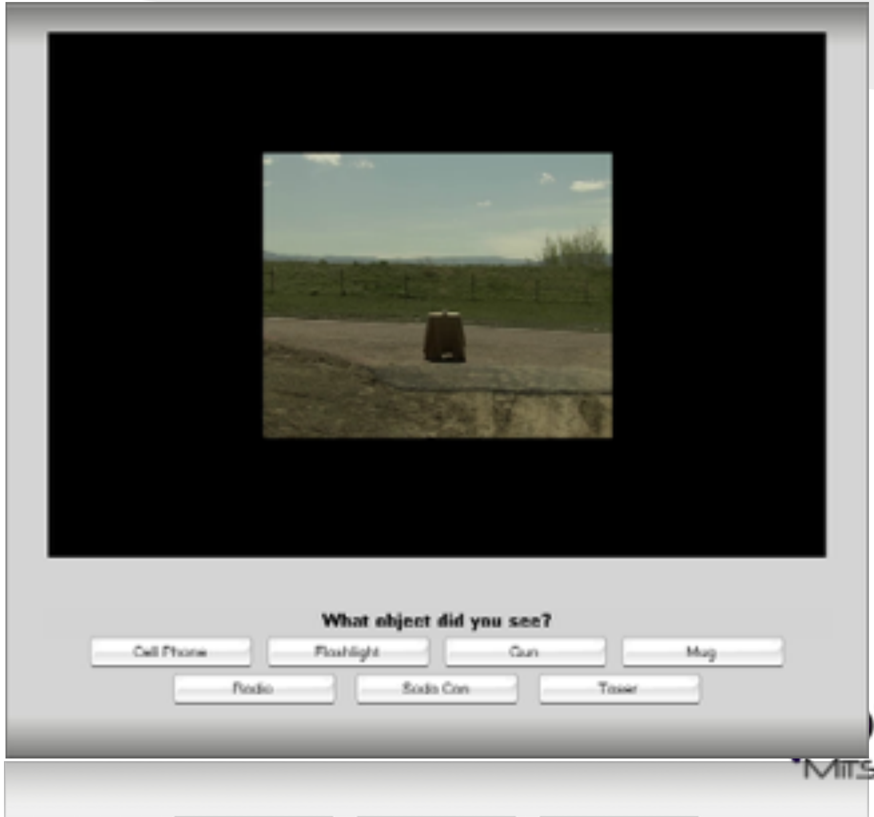


Please answer clip 1 of 20

What was the person holding?

# Multiple-Choice Method (2/4)

- Nothing on impact on choices by buttons':
  - Order
  - Position
- Research showing existence of such impact
- Proposing random sequence of buttons



What object did you see?

# Multiple-Choice Method (3/4)

**Section 6.1 of P.912:**

*The number of choices offered to the viewer will depend on the number of alternative scenes being presented. “Unsure” may be one of the listed choices.*

# Multiple-Choice Method (4/4)

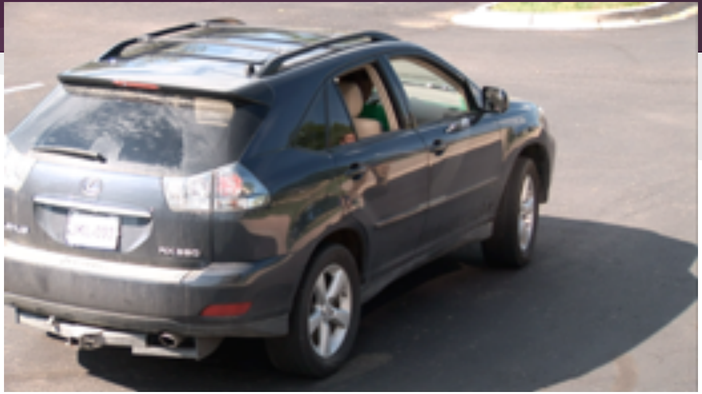
- Subjects observed abusing “Unsure”-like responses:
  - ITU-T P.800 Comparison Category Rating (CCR, Table),“0” (“About the Same”)
  - Similar trend observed independently in TRV study conducted by authors
- P.912 missing clear warning against prudent use of “Unsure” (even encouraging its use)
- Proposing introduction of appropriate entry into P.912

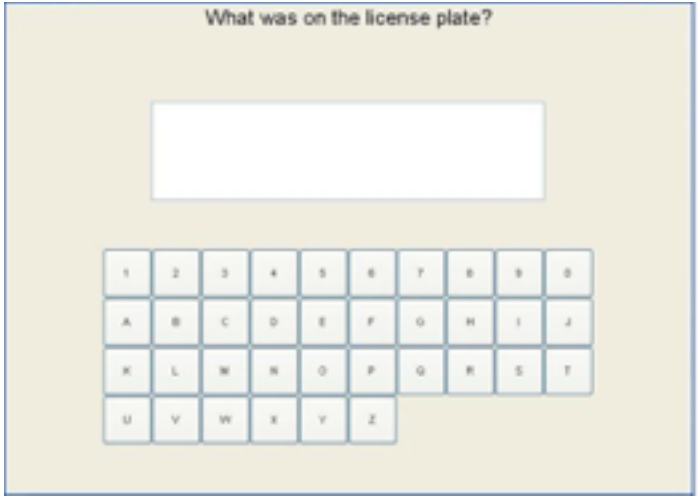
| | |
|---|---|
| 3 | Much Better |
| 2 | Better |
| 1 | Slightly |
| 0 | About the Same |
| -1 | Slightly Worse |
| -2 | Worse |
| -3 | Much Worse |

# Single-Choice Method (1/2)

**Section 6.2 of P.912:**

*If there is a non-ambiguous answer is an identification question, the single answer method may be used. This method is appropriate for alphanumeric character recognition scenarios. A viewer is asked what letter(s) or number(s) was present in a specific area of the video, and the answer can be evaluated as either correct or incorrect.*





What was on the license plate?

# Single-Choice Method (2/2)

- Contrary to P.912, possible to also apply fuzzy logic
- For scenarios with alphanumeric string recognition results, measuring differences between two strings using:
    - Hamming distance (only for strings of same length), or
    - Levensthein distance (generalization of Hamming distance)
- For example, in practice, to consider results containing not more than one error as correct ones
- Even with wrong result of plate recognition, substantially limiting possibility of ultimate mis-recognition, by correlating it with vehicle database containing:
    - Make of vehicle, and
    - Color of vehicle
- Consequently, proposed to expand description of single-choice method.

# Timed Task Method (1/2)

**Section 6.3 of P.912:**

*A viewer may be asked to watch for a particular action or object to be recognized in the video clip. When the viewer perceives that the target has occurred, a timer button can be pushed. In the timed task, the experimenter is able to determine if the time falls within an acceptable time–frame for decision making. These time–frames will be defined by the field in which the video is used, e.g., a person responding to a riot who needs to identify if the crowd has real weapons versus a person who is chasing a car and needs to read the license plate.*

# Timed Task Method (2/2)

- AGH not considering to have expertise/ experience in this area
- But perhaps some people from VQEG do have such expertise/experience
- Any comments to this Section 6.3 of P.912?

# Subjects (1/2)

**Section 7.3 of P.912:**

*Subjects who are experts in the application field of the target video recognition should be used.*

# Subjects (2/2)

- NTIA+AGH experiment on recognizing objects
- Done with expert subjects (LEA officers), **compliance with P. 912, Sec 7.3, but…**
- **Observation:**
  - Experiment repeated also with non-experts, and
  - Got very similar results as long as subjects were paid
- **Conclusion:**
  - Subjects doesn’t need to be experts, if…
  - They are motivated otherwise (e.g.: paid)
  - For some application fields only, impossible for medicine

# Instructing and Training of Subjects (1/3)

**Section 7.4 of P.912:**

*The subject should be given the context of the task before the video clip is played, and told what they are looking for or trying to accomplish. If questions are to be answered about the content of the video, the questions should be posed before the video is shown, so the viewer knows that what the task is.*

**Section 6.2 of P.912:**

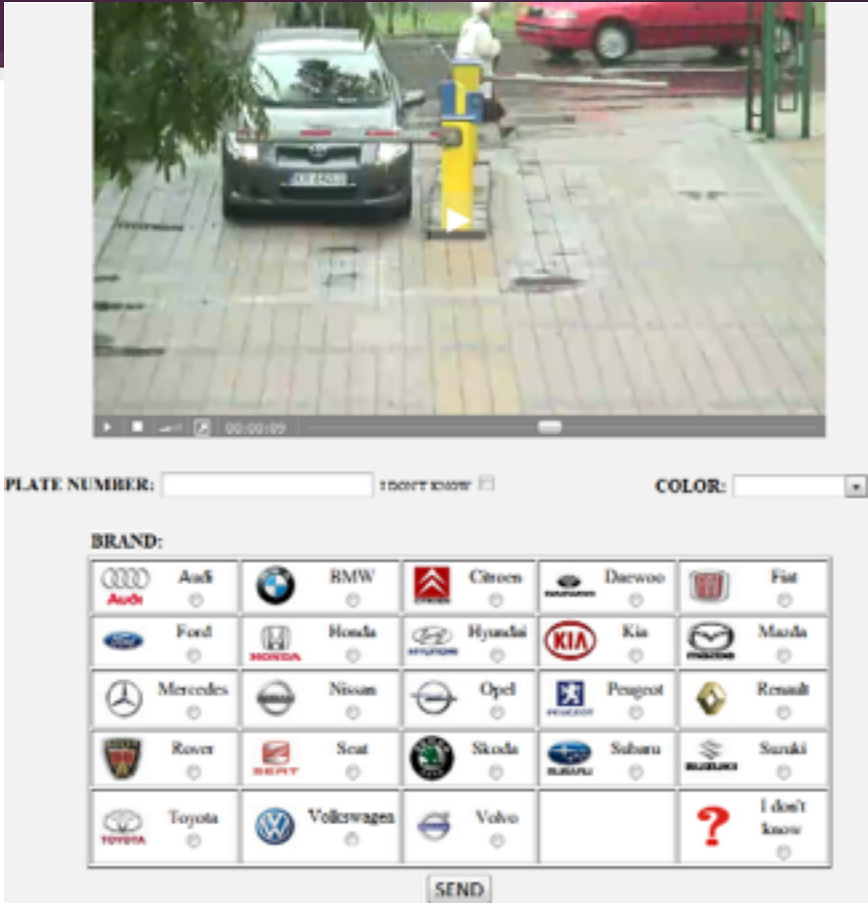
*Care must also be taken to avoid terminology that may differ from participant to participant.*

# Instructing and Training of Subjects (2/3)

- Issues on interacting with subjects not referred in single Section of P.912
- Unnecessary breakdown of topic
- Call for assembling in one (dedicated) Section 7.4 of P.912

# Instructing and Training of Subjects (3/3)

- AGH experiment on recognizing license plates
- Subjects instructed, **compliance with P.912, Sec 7.4, but…**
- Observation:
  - Some subjects recognizing just most obvious characters
  - Others many more of them
- Conclusion:
  - Some subjects assuming to give up on characters difficult to read
  - Others trying hard to read all characters



# Conditions for Testing (1/2)

**Sections: 5, 6, 6.6, 6.7, 7.1, 7.2, 7.3 of P.912:**

*The Experimenter should follow the guidelines outlined in [ITU-T P. 910].*

# Conditions for Testing (2/2)

- P.910 is 1998 and at time of approval P.912 probably most recent on testing conditions to which to refer
- As result, vast majority of tests performed previously under strictly controlled conditions, defined in P.910
- By 2014 P.913 approved largely displacing P.910, including defining smoother requirements for testing
- Calling for introduction of references to P.913, replacing references to P.910

# Statistical Analysis and Reporting (1/2)

**Section 8 of P.912:**

*For single answer conditions, where the answers are correct or incorrect, a statistical metric to determine if the subject is performing above the level of chance for answering correctly should be implemented. "Unsure" answers should be pooled with the incorrect answers.*

*For multiple-choice answers, the probability of an incorrect answer needs to be balanced against the ability to answer the questions correctly. The statistic metric in this situation will require an examination of the stability of the answers within and between subject performance metrics. "Unsure" answers should be pooled with the incorrect answers.*

# Statistical Analysis and Reporting (2/2)

- Statements definitely not taking into account possibility of using more sophisticated statistical techniques
- For statistical analysis of results, authors shown:
  - Possibility of using logistic function
  - Possibility of using Generalized Linear Model (GLZ)
  - Proposals for removing outlier's responses from pool of results - standard procedure in other QoE studies

# *Summary - Standardization*

MITSU

# Summary - Standardization

- Amendments for P.912 being proposed
- Roadmap:
  - Contacting standardization authorities in Poland and USA (done)
  - Consulting amendment at:
    - VQEG meeting, Stockholm, July 2014 (being done)
    - ITU meeting, Geneva, September 2014 (scheduled)
  - Preparing actual edits to P.912 document